Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling

Supplement I - Statistical Methods for Clustering of Gene Expression Data and Validation of Cluster Predictions

Contact:

Yidong Chen[1], Michael Radmacher[2], Richard Simon[2], Amir Ben-Dor[3], Zohar Yakhini[4], Edward Dougherty[5], Michael Bittner[1]

[1]Laboratory of Cancer Genetics, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, [2] Biometric Research Branch, National Cancer Institute, Bethesda, Maryland 20892, [3]Agilent Laboratories, Palo Alto, CA 94304, [4]Chemical and Biological Systems Department, Agilent Laboratories, Haifa 32000 Israel, [5]Department of Electrical Engineering, Texas A & M University, College Station, Texas 77843
(Correspondence: yidong@nhgri.nih.gov)

OVERVIEW:

To fully appreciate the expression patterns derived from large number of cDNA microarrays and their relationship between melanoma tumor samples, several statistical methods were integrated as follows,

1) Multidimensional scaling (MDS) method was employed in order to visualize the similarity between samples, and a hierarchical clustering dendrogram was produced by an implementation of the average-linkage clustering algorithm,
2) The clustering results were further verified by a non-hierarchical algorithm, CAST[1],
3) In order to determine the tightness and the statistical significance of the clusters derived from various methods, two independent approaches were assembled to validate the prediction. One, $WADP_k$ method, is sensitivity analysis of the noise perturbation to the data set. The other one is based on comparing the discrimination power observed for genes in the data to that expected in random data. This is accomplished using TNoM scoring.

4) After confirming the clustering result, each gene was weighted based on their discriminative ability for the clusters derived from previous method.

In the following section, detailed descriptions of the methods listed in Steps 3 to 4 will be presented. For some of the more standard methods, such as MDS, average-linkage methods, and CAST, we refer readers to the literature [1-3]. Since not all genes were readily detectable by the array method, a subset of the total number of surveyed genes was analyzed in all cases. A set of 3613 genes was chosen for analysis. The genes were chosen by an empirically derived set of criteria requiring an average mean intensity above background of the least intense signal (Cy3 or Cy5) across all experiments >2000 arbitrary units, and an average spot size across all experiments of >30 pixels. To avoid distortions of the data resulting from ratios where the signal in one channel is large, and the signal in the other channel is undetectable, ratios higher than 50 or lower than 0.02 were truncated to 50 or 0.02 for these analyses.

Description of the WADP$_k$ method for testing the validity of cluster predictions

Hierarchical clustering of the 31 melanoma samples was performed, resulting in a dendrogram (Fig. 1b). Although the dendrogram gives insights about the similarity and relatedness among samples, it does not indicate robustness to variability associated with the assay sampling, etc. In order to draw valid conclusions about the clustering structure present in the data, it is necessary to investigate how variability affects the results of the cluster analysis. To this end, we developed and implemented a method that determines the reproducibility of given levels of cluster structure within the dendrogram under the condition of added noise. The method is described below.

First, cut the original dendrogram at a height that results in k clusters and let N_k denote the number of clusters containing 2 or more elements. Let M_i represent the number of pairs of elements in the i^{th} of the N_k clusters. Next, perturb the data by adding to every log-ratio of each sample an independent random deviate generated from the $N(0,\_)$ distribution. Cluster the perturbed data and cut the resulting dendrogram at a height that again results in k clusters. For the M_i

pairs of elements in the i^{th} original cluster, record the number of those pairs, D_i that do not remain together in the clustering of the perturbed data. Next, calculate the overall discrepancy rate for the clustering: $(D_1 + D_2 + \ldots + D_{N_k}) / (M_1 + M_2 + \ldots + M_{N_k})$. This overall discrepancy rate is a weighted average of the N_k cluster-specific discrepancy rates (i.e., D_i / M_i, for i = 1, 2, …, N_k), with weights proportional to the number of pairs in individual clusters. Finally, repeat the calculations over many perturbations of the original data set and report the average overall discrepancy rate (termed the Weighted Average Discrepant Pairs for k clusters, or $WADP_k$). The above procedure is repeated for all possible cuts of the original dendrogram and $WADP_k$ is plotted versus k. Minima of the WADP curve are interpreted as indicating reproducible levels of structure.

The parameter σ represents the noise standard deviation inherent to the system. As mentioned above, the noise is composed of—at the least—assay variability and sampling variability. σ is unknown and must be estimated. The method we use for estimating σ is to compute the variance of the log-ratio of each gene across all samples. We then use the median of the empirical distribution of these variances as an estimate of σ^2. It may be more appropriate to use a smaller value (say the tenth percentile of the empirical distribution), if it were believed that a large percentage of genes present on the array were truly differentially expressed within the population of samples hybridized.

Description of the TNoM method for the cluster significance based on random partition.

Threshold number of misclassification, or *TNoM score*, is a simple threshold-based method that uses a given expression level, for a given gene, to predict the cluster label of a given test sample. In the present study, we have 31 samples form 2 groups. Therefore, we can label the samples by l_i , i = 1, …, m, where $l_i \in \{0,1\}$ and m = 31. For the kth gene, let $\langle x_i, l_i \rangle_k$ be its expression pattern (or ratios in this study) and corresponding cluster labels. A threshold function is defined as,

$$f_{h,a}(x) = \begin{cases} a, & \text{if } x < h \\ 1-a & \text{otherwise} \end{cases}$$

where h is a threshold value, and $a \in \{0,1\}$. For a given h and a we can assign the label $f_{h,a}(x_i)$ to the ith sample. The number of misclassifications entailed by this scheme is,

$$e = \sum_{i=1}^{m} \left| l_i - f_{h,a}(x_i) \right|$$

The TNoM score for the *k*th gene, s_k, is defined as the minimum error achieved over all possible choices of *h* and *a*,

$$s_k = \min_{h,a} \left(\sum_{i=1}^{m} \left| l_i - f_{h,a}(x_i) \right| \right)$$

The minimization step is accomplished by exhaustively searching all $2(m+1)$ possibilities.

To examine the significance of groups derived by clustering algorithm, we used three steps. First, We evaluated TNoM scores for all genes found in the data set. Then, the number of genes that have TNoM score less than or equal to *s*, for $s = 0, \ldots, 12$ (where 12 is the maximum misclassifications any classification rule may commit) was listed. Next, we randomly assigned cluster labels to all samples to form two arbitrary groups of 19 and 12 samples. The TNoM score was again evaluated for each gene. A list of the number of genes that have TNoM score less than or equal *s* was similarly obtained. We repeated this process 50 times to observe random fluctuations and their range of scores. Finally, the expected number of genes resulting in *s* or fewer misclassifications under the assumption of perfect random gene expression patterns can be calculated [4]. As expected, the value produced by the 50 random sampling is close to those produced by the theoretical rigorous calculation. The significance of the suggested clusters is reflected in the overabundance of genes with low TNoM scores. More precisely, a meaningful partition will produce far more genes with low TNoM scores than a random one.

Description of the weighting method based on gene's discriminative ability.

The clustering algorithms described in the text produced one tightly bonded cluster of $n_1 = 19$ samples, and we assume the rest of $n_2 = 12$ samples form another cluster. For a given two-cluster setting, a discriminative weight for each gene can be evaluated by,

$$w = d_B \; / \; (k_1 d_{w_1} + k_2 \, d_{w_2} + \alpha)$$

where d_B is the center-to-center distance (between cluster Euclidean distance), d_{w_i} is the average Euclidean distance among all sample pairs, total of t_1 and t_2 sample pairs for cluster 1 and 2, respectively, and $k_1 = t_1/(t_1+t_2)$, and $k_2 = t_2 / (t_1+t_2)$. α is a small constant (0.1 in our study) to prevent zero denominator case. Genes may then be ranked on the basis of *w*. The equation for weight *w* is not only designed to evaluate discriminative ability for single gene, but also capable of evaluate discriminative ability for 2 or more genes together. If you do not assume the second group of samples to be a tight cluster you can drop the d_{w_2} term.

1. Ben-Dor, A., Shamir, R. & Yakhini, Z. *J Comput Biol* **6**, 281-97 (1999).
2. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc Natl Acad Sci U S A* **95**, 14863-8 (1998).
3. Everitt, B. *Cluster Analysis* (Edward Arnold, London, 1993).
4. Ben-Dor, Friedman, Yakhini, submitted for publication.